CoCa: Contrastive Captioners are Image-Text Foundation Models

Year: 2022 | Citations: 1568 | Authors: Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini

---

Abstract

Exploring large-scale pretrained foundation models is of significant interest in computer vision because these models can be quickly transferred to many downstream tasks. This paper presents Contrastive Captioner (CoCa), a minimalist design to pretrain an image-text encoder-decoder foundation model jointly with contrastive loss and captioning loss, thereby subsuming model capabilities from contrastive approaches like CLIP and generative methods like SimVLM. In contrast to standard encoder-decoder transformers where all decoder layers attend to encoder outputs, CoCa omits cross-attention in the first half of decoder layers to encode unimodal text representations, and cascades the remaining decoder layers which cross-attend to the image encoder for multimodal image-text representations. We apply a contrastive loss between unimodal image and text embeddings, in addition to a captioning loss on the multimodal decoder outputs which predicts text tokens autoregressively. By sharing the same computational graph, the two training objectives are computed efficiently with minimal overhead. CoCa is pretrained end-to-end and from scratch on both web-scale alt-text data and annotated images by treating all labels simply as text, seamlessly unifying natural language supervision for representation learning. Empirically, CoCa achieves state-of-the-art performance with zero-shot transfer or minimal task-specific adaptation on a broad range of downstream tasks, spanning visual recognition (ImageNet, Kinetics-400/600/700, Moments-in-Time), crossmodal retrieval (MSCOCO, Flickr30K, MSR-VTT), multimodal understanding (VQA, SNLI-VE, NLVR2), and image captioning (MSCOCO, NoCaps). Notably on ImageNet classification, CoCa obtains 86.3% zero-shot top-1 accuracy, 90.6% with a frozen encoder and learned classification head, and new state-of-the-art 91.0% top-1 accuracy on ImageNet with a finetuned encoder.